LOOKING FOR GOLD IN THE SANDS: STOCK PREDICTION USING FINANCIAL NEWS AND SOCIAL MEDIA

Jiafen Liu, School of Economic Information Engineering, Southwestern University of Finance and Economics, Chengdu, P.R. China, ljfwhy@gmail.com

Qixu Gong, School of Economic Information Engineering, Southwestern University of Finance and Economics, Chengdu, P.R. China, gqx_wolf@qq.com

Tiejun Wang, School of Economic Information Engineering, Southwestern University of Finance and Economics, Chengdu, P.R. China, sohanter@hotmail.com

Wen Zhu, School of Economic Information Engineering, Southwestern University of Finance and Economics, Chengdu, P.R. China, vanechu.cs@gmail.com

Qing Li, School of Economic Information Engineering, Southwestern University of Finance and Economics, Chengdu, P.R. China, Corresponding Author: kooliqing@gmail.com

Abstract

Both traditional finance and behavioral finance theory have reached a consensus that the news media de facto influence stock prices to some extent. There is also evidence that investors are not only subject to the sentiment of related news articles but also the public opinions. The challenge lies on how to quantify such sentimental information to predict the movement of stock market. To measure the sentiments of articles and capture the public mood from postings, we construct and maintain a sentiment dictionary. We utilize both the official information from news articles and user postings in discussion boards to predict firm-specific stock price, and differentiate various types of news articles in the predictive model. Our experiments on CSI 100 stocks during a six week period show a predictive performance in closeness to the actual future stock price is 0.03503 in terms of mean squared error, the same direction of price movement as the future price is 67.6%. Among all seven news topic categories, restructuring news of enterprises has the best predicting performance with direction accuracy of 68.18%.

Keywords: Stock Market, Prediction, Financial News, Social Media, Text Mining.
1 INTRODUCTION

In traditional finance theory, stock price is thought of driven by rational investors to equal the firm's rational present value of expected future cash flows. That’s to say, investors’ decisions in the stock market fully reflect the effects of any information revealed (Fama 1993) which is known as “efficient market hypothesis” (EMH). Lot of research effort has been exerted in predicting the movements of stock market using the quantitative information related with the firm fundamentals (Andersen 2007; Fama 1993; Nartea 2009; Shiller 1981). However, substantial stock market movements cannot be simply captured by the quantitative measures of firms’ fundamentals (Cutler 1989), because our actual market is not efficient as explained by EMH. And the most important is that investors are sentient. In behavioral finance theory, investors are not aggressive in forcing prices to fundamentals as traditional financial theory suggest (Shleifer 1997), and a belief about future cash flows and investment risks is not simply justified by the facts at hand (DeLong 1990). Behavioral financial scholars also believe that investment decisions can be affected by the investors’ emotional impulse. An example is that frustrating news about Steve Jobs' health causes stock prices down of Apple Inc., although its fundamentals are healthy. Even more, recent study shows that the choice of words and tone used by the authors of financial news articles do correlate to measurable stock price movements (Schumaker 2012). Anyway, both sides reach a consensus that the news media does influence asset prices to some extent, but the challenge lies on how to quantify the news articles in the empirical study to investigate their specific impacts on stock market movements.

In recent years, social media has become ubiquitous and important for social networking and content sharing. Therefore, news sources include not only traditional media but also different kinds of social media form such as blogs, micro-blogs, discussion boards, and social news. With social media, investors can timely reach more valuable information than ever before. And social media reflects reaction and emotion of investors traditional more directly than newspaper, television and traditional wire services. So the impact of social media on stock market has been increasing extremely quickly and powerfully.

In this paper, we would like to find answers to following four questions:

- Does the quantitative information in news articles have the ability to affect stock market? If this is true, it proves that the public information events are subject to differential interpretations by investors. This presents profitable trading opportunities for skilled investors, and therefore, the trades of informed investors should be more profitable after the news-releasing day.
- Will incorporating emotion words for quantifying news articles help predict stock price? Given that prior research in predicting stock price has focused mainly on nouns in texts, we throw our sight to emotion words in news articles, so that both important concepts and sentiments of articles can be utilized for stock prediction.
- Will investors react differently to the news articles with different public mood in stock market? We propose to consider public mood while studying the news impact on stock market movements, and expect to verify in our experiments on CSI 100 stocks that the contribution of public mood can not be ignored. If this is true, it provides a concrete evidence to support a critical hypothesis in behavioral finance that investor sentiment affects stock prices.
- Will different topic categories of corporate news cause the difference in the reactions of investors? In prior research, all news articles are treated equally for training the predictive model. Here, we argue that different news categories would affect investors distinctively. For example, the news articles of manager changes can lead to investors’ sensitivity and further overreaction. It would be interesting to differentiate various types of news articles in the prediction.

The rest of this paper is organized as follows. We first briefly describe the related work in Section 2. The design details for our stock price prediction are presented in Section 3. We then test the
performance of our prediction system using real data from the Shanghai Stock Exchange and the Shenzhen Stock Exchange in Section 4. This paper is concluded with speculation on how the current prototype can be further improved in Section 5.

2 RELATED WORK

Observing the fluctuations of stock prices accompanied by the news publishing, some economists have devoted substantial attention to exploring the power of verbal information on stock market. The earliest research is conducted by Cutler, Poterba, and Summers in 1989 (Cutler 1989), and it shows that news articles and stock returns are hardly relevant. But several later findings proved the impacts of news on stock market. For instances, Veronesi found that stock prices overreacted to bad news in good times and underreacted to good news in bad times using a rational expectations equilibrium model of asset prices (Veronesi 1999). Chan empirically examines monthly stock returns following public news and finds that stocks with bad public news display a negative drift for up to 12 months and less drift for stocks with good news (Chan 2003). Tetlock et al. analyze the influence of daily news on stocks and prove the existence of a short-term predictive ability of news for stocks. This study also shows that the fraction of negative words in firm-specific news stories forecasts low firm earnings(Tetlock 2007; Tetlock 2008). Vega further proves the existence of media influence by finding that stocks associated with private information experience low or insignificant drift while stocks associated with public news experience significant drift (Vega 2006).

All these researchers focus on investigating the influence of media on stock market. However, the methods to quantify the influence of news for econometric analysis are questionable. Two general approaches are adopted in these studies. One is treating the number of firm-specific news as a media impactor (Chan 2003; Vega 2006), the other is computing a sentiment indicator based on the percentage of the positive or the negative words in the article (Tetlock 2007; Tetlock 2008). Such methods only partially capture the linguistic power of news, and consequentially weaken or even distort the impact of news on stock market analysis.

While economists are uncovering the relationship of media and stocks, computer scientists are taking a further step to study various media-aware traders to achieve predictable trading returns in stock markets. A pilot study by Wüthrich et al. (Wüthrich 1998) attempts to forecast the one-day trend of five major equity indices in terms of news articles. Later on, several challenging researches are carried out to study the predicable of news on the trends of individual stocks. For instance, Lavernko et al. propose the e-Analyst system to associate stock price trends with the contents of news stories using language modelling techniques (Lavernko 2000). Mittermayer develops the NewsCATS system to predict stock price trends for the time immediately after the publication of press releases (Mittermayer 2006). Fung, Yu and Lam capture the relationships between Reuters Market news and 33 stocks from the Hang Seng Index. One unique contribution of this work is the consideration of the inter-relationship among different stocks, i.e., it selects and assigns relevant news of other similar firms to the target firm while training the predictive model (Fung 2003). Instead of price trend prediction in prior work, Schumaker and Chen propose the AZFinText system to predict the +20 minutes stock price after the news article was released (Schumaker 2009 Quantitative; Schumaker 2009 Textual).

Common strategies of these researches de facto follow a standard paradigm. It starts off with representing the news article as a weighted vector of terms and building a predictive model to capture the relationship between news and stocks. The stock price trends are estimated with this model on the new arrival of a firm-specific news article. Here, quantifying news articles for stock trends analysis is critical and requires a thoughtful design. The common method for textual representation is the bag of words approach. In this approach, an article is represented as an unordered collection of words, disregarding grammar and even word order. Due to the article scaling and noise information, a subset of article terms representing important concepts is adopted to represent the whole text. Schumaker and Chen experiment on various textual representation approaches including bag of words, noun phrases, proper nouns, and name entities, and find out that proper nouns is the most efficient way to quantify
the news articles (Schumaker 2009 Textual). Here, we argue that emotion words are valuable to capture the important topics of news articles since investors are subject to sentiment as proved by the behavioral financial experts. In addition, investors as social creatures may be affected by the behaviors of others. Knowing public feelings or opinions on a stock is helpful to forecast the movements of the stock. More importantly, public mood can strengthen or weaken the impact of news articles on stocks. In particular, investors tend to overreact to negative news in good times and underreact to positive news in bad times (Veronesi 1999). Incorporating public mood can further assist in capturing the relationships between the media and stock market movements. User interactions in social media are good sources to sense public mood. Also, all news articles are treated equally for training the predictive model in research mentioned above. But based on the study of Antweiler and Frank (Antweiler 2012), corporate news can be divided into six main categories: corporate governance, earnings reports, financial issues, operational issues, restructuring issues and legal issues, and the rest fall into "general issues" category. We assume different news topic will affect investors distinctively. Certain types of news may lead to investors’ sensitivity and further overreaction. It would be interesting to differentiate various types of news articles in prediction.

3 SYSTEM DESIGN

The framework of stock price predicting system is sketched in Figure 1. Essentially, it is a news-driven trading system with the consideration of public mood. In this design, it first represents the firm-specific news articles as weighted vectors of terms, and then senses the investors' sentiments on the stock by analyzing recent postings in financial discussion boards. Such information along with stock quotes is fed into the predictive model to capture the relationship between the media and the price of the related stock. The stock selection engine selects a number of stocks for stock price prediction using the well-trained predictive model.

![Figure 1. Design Scheme](image)

In this section, we first describe the way to quantify the news articles to measure their impacts on stocks, and then present the approach to capture public mood on investment atmosphere by digging the postings in financial discussion boards. To assist the sentiment analysis in both media quantification methods, we construct a sentiment words dictionary in financial domain from the Web at the end of this section.

3.1 Representation of news articles

The basic idea to analyze an article is to apply a term vector representation where each element of the vector is a weighted word in the article. Such textual representation is called bag of words model,
which has been used widely in natural language processing and information retrieval. This approach applies almost all the words\(^1\) in the document and generally causes a scaling problem. Essentially, the influence of news articles on stocks originates in two facets:

- **Event:** people tend to adjust their investment strategies if a latest news article conveys some aspects of firms’ fundamentals to enrich their knowledge.
- **Emotion:** the sentimental investors can be affected by the optimistic or pessimistic news.

We model a news article as a weighted term vector \( V \) with a number of nouns and emotion words selected from the article. We believe that the important concepts of firms’ fundamentals can be captured by a group of nouns and the sentiments of news are reflected by the emotion words in the article. Noun detection is a relatively mature technique in natural language processing. Here, we adopt a standard part-of-speech (POS) tagger to extract nouns from the news articles. The challenge of emotion word detection lies on the domain-specific sentiment analysis. There are various of studies focusing on the detection of emotion words in open-domain, but not applicable here. As found by Lougharn and Medonald, there are three-fourths (73.8%) of the negative word counts according to the open-domain emotion word list (Harvard-IV-4) are attributable to words that are typically not negative in a financial context(Loughran 2012). Apparently, building a comprehensive finance-specific emotion word list is of great necessity. We defer the description of how to extract finance-specific emotion words to the section 3.3.

### 3.2 Detection of Public Mood

As we know from psychological study, sentiment, in addition to information, plays an important role in human decision-making (Dolan 2002). In addition, as a social creature, individual is generally affected by others. This could be exaggerated by the stock traders who have higher pressures and expectations (Nofsinger 2005). A good example is the research predicting the stock movement with twitter mood (Bollen 2011).

In this study, we analyze the firm-specific messages in financial discussion boards to capture the public mood on the relevant stock. To this goal, we crawl the postings of the financial discussion boards from Sina.com and EastMoney.com, the most popular stock discussion sites in China. These two websites have more than 20 million independent visiting daily which create a large number of postings and votes. In addition, both sites have a message board for each stock, which matches our needs to capture the investment atmosphere of individual stocks. This data source provides us a solid basis to capture the public mood.

Since the impact of public mood is attenuated but last for several days (Tetlock 2008), the current public mood of a stock should consider the continuing influence of the public feelings during the past several days. Here, we measure the public mood on a stock from two aspects, i.e., optimism and pessimism. Thus, the optimistic mood of a stock, \( s \), on the daily basis is defined as,

\[
M^*_o = \sum_{i=0}^{\tau} \frac{P_i}{L_i} \times e^{i/\beta}
\]

where \( P_i \) is the number of positive words in the firm-specific discussion threads on the \( i \)-th elapsed day, and \( L_i \) is the total number of words in all the firm-specific discussions at that day. \( \tau \) is the number of days that we count for the continuing influence of the public mood, and \( \beta \) is a constant to tune the scale of time attenuation, set to 20 to simulate a month attenuation rate. Similar, the pessimistic mood of a stock \( s \) is

---

\(^1\) A list of so-called stop-words including "the", "of", and "at" are removed because of their semantically empty.
\[ M^e_i = \sum_{i=1}^{\infty} N_i \times e^{-\beta} \]

where \( N_i \) is the number of negative emotion words in the firm-specific discussion threads on the \( i \)-th elapsed day.

### 3.3 Emotion dictionary

Both news representation and public mood detection rely on the technique for judging the sentiment-polarity of a word. A related problem of this issue is sentiment analysis. It maps a given piece of text, such as a document, sentence, or lexicon, to a label drawn from a pre-specified finite set using various supervised or unsupervised machine learning techniques including SVM, Naive Bayes and Maximum Entropy (Pang 2008). Most previous researches focus on the opinion analysis in the open-domain. Here, the challenge lies on the domain-specific sentiment analysis. The general emotion word categorization cannot translate effectively into a discipline with its own dialect (Loughran 2012). Specifically, an emotionless word in the context of finance can express sort of sentiment or a typical emotion word is unemotional in the realm of finance. For instance, the word “bull” originally refers to a male bovine animal but indicates a good earning return in finance domain such as “bull stock”. Some typical emotion words, such as “crude” and “tire” are more likely to identify a specific industry segment than express a negative sentiment in financial events.

To our knowledge, there is no large-scale finance-specific emotion dictionary for our research purpose. In this study, we use an emotion dictionary constructed and maintained by our team. The construction of our emotion dictionary is based on two hypotheses:

- A word is characterized by the company it keeps, i.e., the semantic orientation of a word tends to correspond to the semantic orientation of its neighbors (Turney 2003).
- The positive (negative) orientation of a firm-specific article tends to associate with the upward trend (downward trend) of the relevant stock. This hypothesis is further confirmed by the findings of Tetlock regarding the existence of the interactions between the stock market and the daily bad news contents (Tetlock 2007).

Therefore, we calculate the joint conditional probability of a word with these two hypotheses as follows and select the words with high probabilities.

### 3.4 Stock Trend

To calculate the statistical information for emotion word detection, it requires to associate firm-specific news and discussions with an upward or downward stock trend. Here, the challenge lies on discovering trends from the time series of stock price. Essentially, this is a curve segmentation problem. The basic idea to address this problem is to plot the discrete data into a curve and then segment it into a series of straight lines (Keogh 2001; Lowe 1987; Fung 2003; Rosin 1997). To fully address this problem in the context of stock price, there are several curve patterns to be analyzed properly, i.e. fluctuation and interruption.

- **Fluctuation:** In this study, the purpose of curve segmentation is to find out which articles tend to raise up (or bring down) the relevant stock price. Small fluctuations in a curve should not affect its entire trend. In other word, a small drop (or raise) in a roaring (or slipping) trend should still be treated as a upward (or downward) trend as a whole (Figure 2(a)).
- **Interruption:** Similar to the fluctuation pattern, a significant, but temporary, drop or raise in a curve should not affect its entire trend (Figure 2(b)). Such interruptions are generally caused by noise or errors in the stock transaction data, which should be ignored in our study.
In this article, we propose a segment-and-merge approach for discovering the trend of stock price. It first segments a curve into a number of straight lines and then merges these lines to avoid fluctuation, interruption and over segmentation. This process is repeated until all interruptions are identified and merged.

3.5 Predictive Model

In this study, the function of the predictive model is to capture the relationship between the finance indicators representing the present media and financial situations, and the future stock price. These finance indicators are firm-specific news articles, public mood, and the stock prices at the point of releasing these news articles. The public mood provides a measurement for the recent investment atmosphere, and a firm-specific news article conveys the information of the firm's fundamentals and the attitudes of authorities. These indicators allow us to study whether the facts and emotions in a financial news report coupled with the recent public mood affect the stock market movements.

There are variety of machine learning methods for stock market prediction including Self-organizing Neural Network(SOFNN) (Bollen 2011), language model (Lavenko 2000), Support Vector Machine(SVM) (Mittermayer 2006) and naïve Bayesian (Seo 2002). However, all of these work focus on the directional movements rather than discrete stock prices. In this study, we adopt the extended SVM, i.e., Support Vector Regression(SVR) Model, which applies a regression technique to SVM to support discrete prediction of the stock price (Schumaker 2009 Textual).

4 EXPERIMENTAL SETTINGS

The ultimate goal of this study is to examine the influence of different news articles and public mood on the stock market movements. Here, we target on the stock market in Mainland China. It consists of two independent stock exchanges in the Mainland China, i.e. the Shanghai Stock Exchange (SSE) and the Shenzhen Stock Exchange (SZSE). Prior relevant studies (Schumaker 2009 Quantitative; Schumaker 2009 Textual; Tetlock 2007; Tetlock 2008) are mainly on stock exchanges in Unite States, especially, New York Stock Exchange (NYSE). Due to the lack of market makers in Chinese markets, this study on SSE and SZSE provides a unique insight on the relationship of the media and the stock market without the interference from the trades of market makers.

Three databases are constructed and maintained for our experimental study and further peer study.

Financial News Corpus: This corpus is constructed by querying the Baidu and the Google search engine to download the financial news articles released from Jan.1, 2011 and Dec.31, 2011 with company common names, or abbreviations, or stock number IDs. To remove the noise information, we
keep only the news articles with company name in titles. After removing the HTML tags of each news article, the news body and publication date are stored in this database. This corpus contains 210,086 financial news articles related with 100 companies listed in China Securities Index (CSI 100).²

Financial Discussion Board Corpus: This corpus contains the discussion threads of CSI100 companies during Jan.1, 2011 and Dec.31, 2011 from two premier financial discussion boards in China, i.e., Sina.com and EastMoney.com. Both sites have a message board for each stock in the SSE and the SZSE. More than 2 million active users of these two websites provide a solid base to capture the public mood on individual stocks.

Stock Transaction Data: This corpus contains the high-frequency financial data during Jan.1 2011 and Dec.31, 2011 provided by China Stock Market Database(CSMD). It covers data of all A shares companies listed on the SSE and the SZSE. It provides intraday transaction information including price, volume and time in the second-level.

To gauge how well the proposed approach to capture the movements of stocks, we chose two evaluation metrics as suggested by Schumaker and Chen (Schumaker 2009 Quantitative): closeness and directional accuracy. Directional accuracy measures the upward or downward direction of the predicted stock price compared with the actual direction movement of the stock price. Realized the fact that it may be close in prediction yet predict a wrong movement direction, closeness metric, as a good complement, is proposed to evaluate the difference between the predicted value and the real stock price in terms of Mean Squared Error (MSE). Following the track of previous researches (Gidofalvi 2001; Schumaker 2009 Quantitative; Schumaker 2009 Textual), we adopt 20 minutes as our basic experiment setting to make a fair comparison with previous studies.

To generalize our findings, we select 7-week research period of July 18, 2011 to Sept. 2, 2011 as the basic experimental data from our database without any sudden market or industry-wide price fluctuations³. We remove 11 companies with suspicious abnormal fluctuations during studying period. We split these 30 trading-day data into two parts, the first 5-week data is for training the predictive models, and the fifth week data is for testing.

4.1 News Representation

The influence of news on investors comes from two sources, i.e. events and emotions. Therefore, we extract proper nouns and sentiment words to represent a news article with the assumption that proper nouns reflect firm’s fundamentals and emotion words convey the optimistic or pessimistic attitudes of news.

Here, we adopt the FudanNLP, one of state-of-the-art Chinese lexical analysis systems⁴ as our standard POS tagger to extract seven noun categories including time, place, name, number, entity as our proper noun set from the news articles.

Previous work(Schumaker 2009 Quantitative; Schumaker 2009 Textual; Tetlock 2007; Tetlock 2008) rely on Harvard Psychosociological Dictionary (Harvard-IV-4) to capture the tone of an article. Here, we argue that the emotion word list developed for psychology and sociology cannot function well in the realm of finance. In fact, Loughran and MacDonald (Loughran 2012) point out that Harvard-IV-4 list substantially misclassifies words when gauging tone in financial applications. In this article, we construct a finance-specific sentiment word list and represent the emotions of financial articles with

---

² CSI 100 consists of the largest 100 stocks in the mainland China. CSI 100 aims to comprehensively reflect the price fluctuation and performance of the large and influential companies in Shanghai and Shenzhen securities markets.

³ To construct a comprehensive finance-specific sentiment word dictionary, we apply the entire 2011-year data of news articles and postings.

⁴ It is developed by Fudan University, and accessible at http://code.google.com/p/fudannlp
these sentiment words. To our knowledge, this is the largest and comprehensive Chinese sentiment word list in financial domain.

In our experiments, we study several ways to represent news articles. The predictive performance of these representations is shown in Table 1. $N_{\text{properNoun}}$ denotes that the news article is represented by a number of weighted proper nouns. $E_{\text{Harvard}}$ denotes that the article is represented by a number of weighted emotion words from Harvard-IV-4 list. $E_{\text{Finance-specific}}$ denotes that the article is represented by a number of weighted emotion words from our finance-specific emotion word list. As we can see, the representation based on the finance-specific emotion words can greatly improve the predictive performance. Integrating the finance-specific emotion words into the proper noun representation can further improve the predictive performance.

<table>
<thead>
<tr>
<th>Method</th>
<th>MSE</th>
<th>Direction</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N_{\text{properNoun}}$</td>
<td>0.0542</td>
<td>52.4%</td>
</tr>
<tr>
<td>$E_{\text{Harvard}}$</td>
<td>0.0472</td>
<td>54.2%</td>
</tr>
<tr>
<td>$E_{\text{Finance-specific}}$</td>
<td>0.0402</td>
<td>61.4%</td>
</tr>
<tr>
<td>$E_{\text{Finance-specific}} + N_{\text{properNoun}}$</td>
<td>0.0382</td>
<td>64.9%</td>
</tr>
</tbody>
</table>

*Table 1: News Representation*

### 4.2 Public Mood

In behavioral finance theory, investors are subject to sentiment in their decision-making. Information and public mood are two important sources to affect the feelings of the investors. In this study, we analyse the firm-specific messages in two premier financial discussion boards from Sina.com and EastMoney.com to capture the public mood on individual stocks. We carry out a series of experiments to study the influences of public mood ($M_+^s$ and $M_-^s$) on a stock. In particular, we create three models to compare the influences of public mood with different attitudes as shown in Figure 4. For each stock, Model 1 takes its present stock price, the term vectors of relevant news articles, and the optimistic public mood on this stock ($M_+^s$) as the input of the predictive model. Model 2 takes the present stock price, the term vectors of news articles, and the pessimistic public mood ($M_-^s$) as the input. In model 3, the input are the present stock price, the term vectors of news articles, $M_+^s$ and $M_-^s$. The output of these models is +20 minute stock price.

![MSE](image1.png)  
![Directional Accuracy](image2.png)

*Figure 3. Public Mood*

From Figure 3, the pessimistic public mood has a significant contribution to predict the stock movements and its impact can be reach up to 5 days. Comparing with pessimism attitude, the
optimistic public mood has a limited power to sense the movement of stocks. These findings cooperate nicely with the prior study that the fraction of negative words in firm-specific news stories forecasts low firm earnings (Tetlock 2008). The joint influence of pessimism and optimism in public mood is observable.

4.3 Topic category

In prior research, all news articles are treated equally for training the predictive model. Based on the study of Antweiler and Frank (Antweiler 2012), the topic categories of corporate news may affect investors distinctively. For example, we assume restructuring news will draw investors’ more attention and cause further reaction according to investor’s experience.

In Financial News Corpus, we select 15,759 news articles from August 1, 2011 to August 31, 2011, and tag them with seven labels, i.e., “Corporate Governance”, “Earnings Reports”, “Financial Issues”, “General Issues”, “Legal Issues”, “Operational Issues”, and, “Restructuring Issues”. Then we use SVM to classify the rest news articles in financial news corpus and get satisfying classification result with density 0.999729 and F-measure 0.74. Then we choose 6 weeks research period of July 18, 2011 to August 26, 2011 for training the predictive models, one-week data from August 29, 2011 to September 2, 2011 for testing. Articles gathered during this period were restricted to occur between office time of Chinese stock market, and two articles on the same company cannot exist within 20 min. The experiment data and result are shown as in Table 2.

<table>
<thead>
<tr>
<th>Category</th>
<th>Training</th>
<th>Testing</th>
<th>MSE</th>
<th>Direction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corporate Governance</td>
<td>59</td>
<td>3</td>
<td>0.0273</td>
<td>0.0000</td>
</tr>
<tr>
<td>Earnings Reports</td>
<td>1756</td>
<td>342</td>
<td>2.5699</td>
<td>53.6443</td>
</tr>
<tr>
<td>Financial Issues</td>
<td>358</td>
<td>75</td>
<td>1.1324</td>
<td>44.7368</td>
</tr>
<tr>
<td>General Issues</td>
<td>784</td>
<td>112</td>
<td>0.1416</td>
<td>47.7876</td>
</tr>
<tr>
<td>Legal Issues</td>
<td>142</td>
<td>23</td>
<td>0.4455</td>
<td>29.1667</td>
</tr>
<tr>
<td>Operational Issues</td>
<td>1616</td>
<td>351</td>
<td>1.3710</td>
<td>54.2614</td>
</tr>
<tr>
<td>Restructuring Issues</td>
<td>157</td>
<td>21</td>
<td>0.1117</td>
<td>68.1818</td>
</tr>
</tbody>
</table>

Table 2. Topic Category

First of all, we notice that category of corporate governance does not obtain any valuable result for its poor sample size, so this category should be ignored. It is observable that restructuring issues and operational issues have better predicting performance in direction accuracy than other categories. Restructuring issues is also of the minimal MSE metric, but closeness results of other categories indicate large deviation in predicting. Then we remove 7 stocks in wine industry, such as GuizhouMaotai(600519), whose stock price is extremely high to disturb the price prediction accuracy. As we can see in Table 3, closeness of prediction using earning reports and operational issues is highly improved, and direction accuracy using general issues and operational issues is slightly lifted.

<table>
<thead>
<tr>
<th>Category</th>
<th>MSE</th>
<th>Direction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Earnings Reports</td>
<td>0.1151</td>
<td>52.5680</td>
</tr>
<tr>
<td>Financial Issues</td>
<td>1.1324</td>
<td>44.7368</td>
</tr>
<tr>
<td>General Issues</td>
<td>0.1416</td>
<td>48.2143</td>
</tr>
<tr>
<td>Legal Issues</td>
<td>0.4455</td>
<td>29.1667</td>
</tr>
<tr>
<td>Operational Issues</td>
<td>0.3514</td>
<td>54.8673</td>
</tr>
<tr>
<td>Restructuring Issues</td>
<td>0.1117</td>
<td>68.1818</td>
</tr>
</tbody>
</table>

Table 3. Performance after Removing
We believe that the stock price predicting performance will be improved after we expand the scale of training data set. Since we have known different news categories affect investors distinctively, we could compute the expected price of stock with more accuracy using weighted news topic.

5 CONCLUSION AND FUTURE WORK

There are several interesting findings in our research. The first is that the media influence of financial news on stocks is existed and can be quantified using text mining techniques. In particular, the fundamental information of a firm-specific news article can enrich the knowledge of investors and affect their trading activities. Meanwhile, the sentiment of news articles may lead to the emotion fluctuations of investors and interfere their decision-making. Both hypotheses are nicely supported by our experimental results that representing news articles with a number of proper nouns and emotion words provides the optimal way to quantify the news for stock prediction. Second, as social creatures, the behaviours of investors are interactive. The public mood of others on a stock can affect individual investment decisions. Based on our experimental results, the influence of public mood is continuously increasing within 5 days and become weaker. With the popularity of Web 2.0, there are various types of social media for readers to express their opinions including Blogs, Micro-blogs, and social news. An investigation into these new sources to capture the public mood would be essential.

We still have a long but exciting journey ahead of us. There are so many wonderful things for us to explore in the ocean of information and data, and so many interesting questions for us to answer. Our future work would cover the exploration of other factors including company size, industry category and media exposure.

6 ACKNOWLEDGMENT

This work has been supported by the National Natural Science Foundation of China (NSFC Grant No. 60803106, 61170133, 60903201, 91218301), the Fok Ying-Tong Education Foundation, China (Grant No. 121068), the Program for Key Innovation Research Team in Sichuan Province (2011JTD0028) and Key Innovation Research Team in SWUFE.

REFERENCES


